Black-Box Adversarial Defense via Anomaly Detection

Xing Binbin

Аспирант

Факультет ВМК МГУ имени М.В. Ломоносова, Москва, Россия

E-mail: xingbinbin@cs.msu.ru

Научный руководитель *— Гамаюнов Денис Юрьевич*

Deep learning models exhibit significant vulnerability to adversarial examples, posing security risks in critical applications like autonomous driving and medical diagnosis. Black-box attacks are particularly concerning as they reflect realistic scenarios where attackers have only API-level access. This study systematically compares three anomaly detection approaches for defending against such attacks, based on the hypothesis that adversarial examples, while able to deceive classifiers, exhibit feature distributions that deviate sufficiently from normal samples to be detectable.

We evaluate three representative methods:

Isolation Forest [1] constructs random partition trees by recursively splitting features. Anomalies are identified through shorter path lengths as they require fewer partitions to isolate.

One-Class SVM [2] learns a minimal hypersphere enclosing normal samples in kernel space. Samples outside this boundary are flagged as anomalies based on their distance from the center.

Autoencoder [3] detects anomalies through reconstruction error. The architecture compresses inputs into latent representations then reconstructs them, with higher errors indicating potential adversarial samples.

Experimental Results:Our evaluation on MNIST dataset yields several key findings:

Computational Efficiency: One-Class SVM achieves the fastest inference at 13.63s/sample.

Memory Usage: One-Class SVM requires only 8.7MB vs Autoencoder's 22.4MB.

Detection Performance: All methods show comparable accuracy under current evaluation setup.

As shown in Рис. 1,, traditional methods demonstrate clear efficiency advantages. This makes them particularly suitable for resource-constrained real-time applications. However, deep learning approaches show promise in handling more complex data distributions.

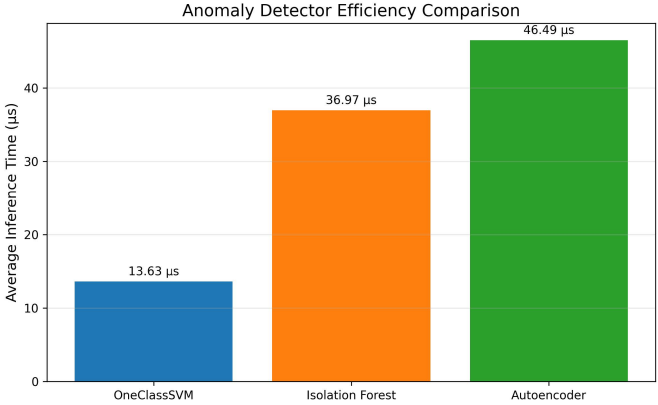

Рис. 1: Computational efficiency and memory usage comparison.

Литература

1. Szegedy C., Zaremba W., Sutskever I. et al. Intriguing properties of neural networks // arXiv preprint arXiv:1312.6199. 2013.
2. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. 2014.
3. Papernot N., McDaniel P., Jha S. et al. The limitations of deep learning in adversarial settings // 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016. P. 372–387.